Sztuka Leczenia
2023, tom 38, nr 2, ss. 27–36
https://doi.org/10.4467/18982026SZL.23.012.18588
https://www.ejournals.eu/SL/

Krzysztof Mudyń

Akademia Ignatianum w Krakowie, Instytut Psychologii

Ekspansja sztucznej inteligencji a problem wartości

The expansion of artificial intelligence and the problem of values

STRESZCZENIE

Autor poszukuje związków między ekspansją sztucznej inteligencji (SI) a problematyką wartości. Szczególną uwagę zwraca na trudności związane z „uzgadnianiem wartości" w ramach interakcji „algorytm vs człowiek". Przybliża problemy związane z uwzględnianiem ludzkich wartości w projektowaniu złożonych algorytmów. Wynikają one z faktu, że preferowane wartości nie są spójne i mają charakter kontekstowy (a więc zmienny). Są też zależne od uwarunkowań kulturowych i różnic indywidualnych. Szczególnie trudne do przewidzenia są tzw. wartości sentymentalne. Wszystko to sprawia, że jednoznaczne zdefiniowanie respektowanych przez algorytm wartości jest prawie niemożliwe. Obecnie podejmowane są próby uwzględniania *emotional computing* w projektowaniu sztucznych systemów, co zdaniem wielu badaczy może okazać się przełomem w rozwoju SI. Podjęto już zaawansowane próby modelowania jednego z aspektów inteligencji emocjonalnej, jakim jest rozpoznawanie cudzych stanów emocjonalnych w oparciu o analizę mimiki twarzy. Zdaniem autora sukcesy w zakresie *artificial emotional intelligence* powinny raczej martwić niż cieszyć użytkowników Sieci. Poskutkują bowiem większą kontrolą ze strony używających ich instytucji, a w konsekwencji do dalszym ograniczaniem wolności osobistej indywidualnych użytkowników. Mudyń sugeruje, że ekspansja technologii cyfrowej (wbrew początkowym nadziejom) przyczynia się do zwiększonej centralizacji władzy i nierówności społeczno-ekonomicznych. Mówiąc słowami Norberta Wienera (1950), ojca cybernetyki, rozwój technologii cyfrowej przyczynia się do „the human use of human beings".

Słowa kluczowe: sztuczna inteligencja, nieokreśloność i niewspółmierność wartości, „sztuczna inteligencja emocjonalna", wygoda vs wolność osobista

ABSTRACT

The author analyzes relation between the expansion of artificial intelligence (AI) and the issue of values. In particular, he points out the difficulties associated with "agreeing on values" in an algorithm vs during a human interaction. The article highlights the obstacles of taking human values into account while designing complex algorithms, which result from the fact that the

Adres do korespondencji / Address for correspondence: km.krzysztof.mudyn@gmail.com, krzysztof.mudyn@ignatianum.edu.pl
ORCID: https://orcid.org/0000-0001-6177-7241

preferred values are inconsistent, contextual and therefore variable. The values depend on cultural conditions and individual differences as well. In addition, the sentimental values are also difficult to predict and take into account. All this makes it almost impossible to unambiguously define the values to be respected by the algorithm. Currently, an attempt is being made to include "emotional computing" into a design of artificial systems, which, according to many researchers, may turn out to be a breakthrough in the development of AI. There are already advanced attempts being made to model one of the aspects of emotional intelligence, which is to recognise other people's emotional states based on the analysis of their facial expressions. According to the author, developments in the field of artificial emotional intelligence should rather worry than satisfy the users of the internet. They will contribute to greater control exercised by the institutions that use them, and consequently to further limitation of personal freedom of the individual users. The author suggests that the expansion of digital technology (contrary to the initial hopes) contributes to increased centralization of power and socio-economic inequalities. In the words of Norbert Wiener (1950), the development of digital technology contributes to "the human use of human beings".

Keywords: artificial intelligence, indefiniteness and incommensurability of values, "artificial emotional intelligence", comfort vs personal freedom

Wprowadzenie

Rewolucja cyfrowa, z konsekwencjami której stykamy się na co dzień, stopniowo, acz radykalnie zmienia nasz styl życia, nawyki i preferowane wartości. Trudno znaleźć taką dziedzinę życia, na której technologia cyfrowa nie odcisnęła jeszcze swego piętna. Oprócz zdalnych, niejako eksternistycznych kontaktów interpersonalnych, mamy do czynienia ze zdalną edukacja, pracą i szerokim wachlarzem wirtualnych rozrywek. Technologia cyfrowa wpłynęła też na sposób zagospodarowywania wolnego czasu, szczególnie przez najmłodszych użytkowników. Trwają ożywione dyskusje nad przyszłością sztucznej inteligencji (SI). Stanowiska w tej kwestii są bardzo spolaryzowane, poczynając od entuzjastycznych zwolenników, którzy w jej rozwoju są skłonni upatrywać zbawienia ludzkości, kończąc na nieufnych sceptykach, którzy formułują kasandryczne przepowiednie.

W prezentowanym tekście nie zamierzam rozstrzygać tego dylematu. Skupię się na obszarach, w których SI styka się z problematyką wartości. Niezależnie od przebiegu toczących się dyskusji nasz wpływ na kierunek ewolucji technologicznej (podobnie jak w przypadku ewolucji biologicznej) jest minimalny, niemal zerowy. Lawina ruszyła kilka dekad temu i toczy się coraz szybciej. Pytanie brzmi raczej: „Co zrobić, aby całkiem nas nie przygniotła?". Z psychologicznego punktu widzenia nietrudno dostrzec sporo niepokojących konsekwencji i ciemniejszych stron tej ewolucji. Technologia cyfrowa kusi nas różnymi wersjami teleobecności, okradając nas tym samym z pełnej obecności w fizycznych i społecznych kontekstach, w których porzucamy nasze ciała. Można powiedzieć, że zmienił się nasz „sposób bycia w świecie", że mamy do czynienia z „dekontekstualizacją istnienia" (Mudyń, 2010). Oznacza to, że raz po raz nasza dominująca w danym czasie aktywność mentalna rozgrywa się poza fizycznym i społecznym kontekstem, w którym znajdują się nasze ciała. Prawdopodobnie prowadzi to do modyfikacji systemu wartości i nie tylko. Wprawdzie zmiany międzypokoleniowe trudno jest badać, gdyż interferują ze zmianami związanymi z wiekiem osób badanych, lecz niektóre panelowe porównania dostarczają pośrednio argumentów, że tak się dzieje. Badania prowadzone na populacji holenderskiej (Leijen, Herk, Bardi, 2022) pokazują, że pokolenie millenialsów różni się od trzech pozostałych, tj. *silent generation*, *baby boomers* i *generation X*, dynamiką zmian zaobserwowanych w trakcie 12-letnich badań.

Dla porządku należy jednak wspomnieć o najmniej kontrowersyjnych i korzystnych aspektach cyfryzacji. Niewątpliwie rozwój SI wniósł wiele dobrego do nauk medycznych, szczególnie tych, gdzie medycyna styka się inżynierią – chirurgii, protetyki, wczesnej diagnozy czy zdalnego monitoringu procesów fizjologicznych. Zauważmy też, że rozwój telekomunikacji cyfrowej, m.in. za

sprawą aplikacji typu Microsoft Teams, Zoom itp. okazał się bardzo pomocny w przetrwaniu wymuszonej izolacji podczas pandemii Sars-CoV-2.

Cel pracy

Generalnie jesteśmy skłonni wierzyć, że SI (przynajmniej póki co) służy ludziom. Znaczyłoby to, że pomaga im realizować ich cele i wartości. Przy bliższym spojrzeniu pojawia się jednak wiele pytań i wątpliwości. Którym ludziom? W jakim stopniu skuteczność realizacji instrumentalnych celów zmienia lub modyfikuje nasze preferencje? W nawiązaniu do terminologii Miltona Rokeacha (1973) można by zapytać, na ile „wartości instrumentalne" wpływają na doświadczanie i realizowanie „wartości ostatecznych".

Coraz wyraźniej dostrzegamy dylemat „bezpieczeństwo vs prawa jednostki". Mniej oczywistym natomiast jest „wygoda vs swoboda wyboru". Coraz więcej uwagi projektantów przyciąga też problematyka „uzgadniania wartości", czyli uczenia sztucznych systemów ludzkich wartości (oraz psychologii) i *vice versa*. Ludzkie preferencje (wartości) są trudne do zaimplementowania, gdyż są niespójne, zależne od kontekstu i nie całkiem jawne. Krótko mówiąc: są trudno definiowalne. Obecnie trwają też badania nad tzw. sztuczną inteligencją emocjonalną. Pojawia się pytanie: „Czy dysponenci naturalnej inteligencji emocjonalnej powinni podzielać entuzjazm jej projektantów i producentów?". Zasygnalizowane tu kwestie będą rozwijane i uzasadniane w kolejnych partiach tekstu.

Czy SI jest inteligentna?

Póki co algorytmy sztucznej inteligencji, mimo spektakularnych często wyników, nie zasługują na miano inteligentnych. Pomijając literaturę science fiction, mamy do czynienia *de facto* z wyrafinowaną symulacją rozwiązań i rozstrzygnięć, które – gdyby były przejawem aktywności człowieka (tu nawiązuję do genezy tego terminu) lub innych organizmów – bylibyśmy skłonni uznać za przejawy inteligencji. Zauważmy, że nawet najbardziej złożone algorytmy, z tzw. autonomicznymi pojazdami, nie są autonomiczne, gdyż nie potrafią samodzielnie podtrzymywać swojego istnienia poprzez poszukiwanie odpowiednich zasobów energetycznych. Nie posiadając instynktu samozachowawczego, nie są również w stanie przeciwdziałać destrukcyjnym oddziaływaniom środowiska i modyfikować swojej struktury stosownie do zmian otoczenia. Krótko mówiąc, zasadniczo nie posiadają one zdolności adaptacyjnych. Z tego punktu widzenia rośliny są inteligentniejsze od SI, ponieważ w sposób aktywny i skuteczny potrafią adaptować się do zmian zachodzących w środowisku (Mudyń, 2022). Algorytmy nie potrafią też się rozmnażać, czyli samodzielnie powielać swojej struktury. A zatem wygląda na to, że jakościowa zmiana w rozwoju SI będzie uwarunkowana postępami w konstruowaniu sztucznego życia.

Algorytmy potrafią jednak bardzo szybko się uczyć (m.in. za pośrednictwem Sieci) i równie szybko dzielić się nabytymi umiejętnościami z innymi wirtualnymi maszynami. Toby Walsh (2018, s. 153) przybliża tę kwestię, pisząc – „Gdy jeden samochód Tesla nauczy się rozpoznawać wózek jadący na zakupy i unikać go, możemy załadować nowy kod do całej floty Tesli na całym świecie". I dodaje – „Jest to tak ważna sprawa, że wymyśliłem na nią nową nazwę *co-learning*".

Zważywszy, że w stałych i dobrze zdefiniowanych okolicznościach algorytmy potrafią „zachowywać się" dość adekwatnie, chętnie i mimowolnie je antropomorfizujemy (Mudyń, 2012, 2014), co może być przyczyną nieadekwatnych oczekiwań i kosztownych błędów. Nawet w przypadku, gdy dzwoniąc do jakiejś instytucji zostajemy poinformowani (ludzkim głosem, sic!), że rozmawiamy ze sztuczną inteligencją, to w trakcie konwersacji łatwo zapominamy, że mamy do czynienia z SI i próbujemy z nią negocjować lub się wykłócać. Algorytmy potrafią też popełniać takie błędy, które z inteligencją nie mają nic wspólnego i których się nie spodziewamy. Gerd Gigerenzer (2022) przytacza przykład, który zyskał własną nazwę, a mianowicie *Russian tank fallacy* („błąd rosyjskiego czołgu"). Otóż sztuczny system próbowano nauczyć odróżniania rosyjskich czołgów od amerykańskich. W warunkach laboratoryjnych algorytm uzyskiwał znakomite wyniki, tj. niemal bezbłędnie identyfikował na zdjęciach rosyjskie czołgi. Natomiast w nowych, poza-laboratoryjnych warunkach, zupełnie utracił tę niejako wyuczoną umiejętność.

Powód okazał się banalnie prosty – na wszystkich zdjęciach rosyjskich czołgów, wykorzystywanych w trakcie treningu, był śnieg. I ta okoliczność, nie bez racji, została potraktowana jako najbardziej diagnostyczna wskazówka.

Analogiczna sytuacja zaistniała w przypadku algorytmu, który został pomyślany jako cyfrowe wsparcie przy ocenie rokowań w odniesieniu do pacjentów pulmonologicznych. W nowojorskim szpitalu, gdzie był projektowany i testowany, wykazywał się bardzo wysoką trafnością. Jednak zastosowany w innych szpitalach utracił tę umiejętność. Okazało się, że zdjęcia radiologiczne wykorzystywane w procesie uczenia algorytmu różniły się pod względem technicznym – zdjęcia pacjentów w zaawansowanym stadium choroby (czyli pacjentów leżących) wykonywano przy użyciu innego, przenośnego urządzenia, co stanowiło dla algorytmu bardzo dobrą wskazówkę diagnostyczną.

Niedookreśloność i niewspółmierność preferowanych wartości

Problemy nie tylko natury technicznej, lecz także teoretycznej i etycznej ujawniają się – co zrozumiałe – w trakcie interakcji konkretnych ludzi z różnymi aplikacjami SI, które – z założenia – powinny im służyć i pomagać. Natychmiast pojawia się pytanie: „Jakim ludziom?". Użytkownikom czy producentom? Jeśli wytwory te mają służyć użytkownikom, to z konieczności muszą korespondować z ich potrzebami oraz uwzględniać najbardziej uniwersalne wartości. Kłopot polega te na tym, że – jak wiadomo – algorytmy SI bardzo dobrze radzą sobie tylko w stabilnym i dobrze zdefiniowanym otoczeniu. W przypadku jednoznacznie zdefiniowanych gier, takich jak szachy, nawet najwybitniejsi szachiści nie są w stanie im sprostać. Ludzkich wartości jednak nie da się jednoznacznie zdefiniować i uwzględnić w złożonym algorytmie.

Dodajmy, że preferowane wartości nie są abstrakcjami, nie są tym, do czego chętnie odwołują się politycy w sytuacjach publicznych. Są raczej tym, co sprawia, że dokonujemy takich, a nie innych wyborów, w sytuacjach, gdzie taki wybór potencjalnie jest możliwy. Czyli wówczas, gdy nie jesteśmy pod presją zagrożenia życia, a w posiadanym repertuarze zachowań dysponujemy więcej niż jedną reakcją. Upraszczając – wartości nie są tym, co się deklaruje, lecz raczej tym, co się realizuje, co ujawnia się w zachowaniu i spontanicznie podejmowanych decyzjach. W konkretnych sytuacjach kierujemy się raczej wartościami niejawnymi niż deklarowanymi. Uwzględnianie ludzkich wartości w projektowaniu sztucznych systemów, z robotami włącznie, jest bardzo trudne. Powodów jest kilka.

Po pierwsze, preferowane wartości są zależne od kontekstu. Innymi słowy, nasze preferencje są sytuacyjnie zmienne. Upraszczając, zdrowie staje się wartością nadrzędną wprzypadku zagrożenia własnego życia, w kontekście niekorzystnej diagnozy, lecz w wielu innych sytuacjach tak nie jest.

Po drugie, akceptowane wartości nie są wystarczająco spójne, a nawet pozostają w stanie potencjalnego, cyklicznie powracającego konfliktu. Trudno zatem przewidzieć, która opcja uzyska przewagę w konkretnej sytuacji i ujawni się w postaci wybranej reakcji.

Po trzecie, akceptowane wartości są w dużym stopniu niewspółmierne (niekompatybilne). Najłatwiej to zauważyć przy zderzeniu wartości utylitarnych (skuteczność, opłacalność itp.) z etycznymi, np. uczciwością czy prawdomównością. Co więcej, ta niewspółmierność odnosi się także do samych wartości etycznych. Dla przykładu – kiedy należy być bardziej empatycznym niż prawdomównym, a kiedy odwrotnie? W pewnym sensie dotyczy to także wartości teologicznych – jak pogodzić boską sprawiedliwość z innym atrybutem boskości, tj. miłosierdziem?

Po czwarte, oprócz różnic indywidualnych istnieją też istotne różnice kulturowe (por. kultury indywidualistyczne vs kolektywistyczne). Ponadto w różnych kulturach występują różnego typu tabu zwykle nie rozpoznawane przez przedstawicieli innych kręgów kulturowych. Postulat Horacego *Dulce et decorum est pro patria mori* bardziej przekonująco brzmi zapewne dla przedstawicieli kultur kolektywistycznych niż indywidualistycznych.

Po piąte, istnieją też tzw. wartości sentymentalne, których nie można zaliczyć ani do utylitarnych, ani (na ogół) do moralnych, a są związane zazwyczaj z cennymi pamiątkami i symbolami, które dla danej osoby lub grupy mają szczególną wartość emocjonalną. Tytułem ilustracji odwołam

się do przerysowanego przykładu podanego przez Stuarta Russella (2020). Otóż wyobraźmy sobie robota zaprojektowanego do przygotowywania zdrowych i zasobnych w białko zwierzęce posiłków, który serwuje nam na kolację naszego kota, nie wiedząc, że „wartość sentymentalna tego zwierzaka przewyższa jego wartość odżywczą" (Russell, 2020, s. 50).

Problem polega na tym, że sztuczny system, zaprojektowany do optymalnej realizacji specyficznego celu, może go realizować nadspodziewanie „dobrze", tyle że z pominięciem szerszego kontekstu i ludzkich preferencji, które *explicite* nie zostały uwzględnione w projekcie. Mówiąc słowami Russella – „To, że przydzielimy maszynie jakiś ustalony cel, nie oznacza, że automatycznie rozpozna ona znaczenie, jakie mają dla nas rzeczy niebędące częścią celu. Maksymalizacja celu może równie dobrze ściągnąć na ludzi kłopoty, ale maszyna z definicji nie uzna ich za kłopotliwe" (Russell, 2020, s. 49). Należy podkreślić, że kłopot polega nie tylko na tym, że roztargnieni projektanci coś przeoczą, nie uwzględniając w procedurze optymalizacji dodatkowych zmiennych, które dla ludzi mają istotne znaczenie. Problem jest szerszy, a wiąże się z tym, że my, ludzie, słabo orientujemy się we własnych preferencjach. Wiele badań z zakresu psychologii społecznej i poznawczej (por. Mayers, Twenge, 2018; Nisbett, 2016, s. 65–81) dostarcza przekonujących przykładów, jak bardzo potrafimy się mylić, próbując określić motywy własnego postępowania lub czynniki zewnętrzne, które wpłynęły na naszą decyzję. Poza tym, jak wspomniano wcześniej, nasze preferencje są zazwyczaj bardzo niespójne i zależne od kontekstu.

Inne aspekty wartości w kontekście SI

Jak wiadomo, sztuczne systemy cyfrowe w porównaniu z ludźmi są bezkonkurencyjne w zakresie szybkości przetwarzania informacji i uczenia się optymalnych reakcji w przypadku precyzyjnie i jednoznacznie zdefiniowanych zadań. Ludzkie zachowania i preferencje z pewnością do takich nie należą. Zważywszy, że racją bytu sztucznych systemów jest wchodzenie w interakcje z systemami żywymi, a zwłaszcza z ludźmi, coraz więcej badań dotyczy tej problematyki. Pojawił się nowy termin dotyczący kooperacji ludzi z „inteligentnymi" maszynami, czyli tzw. dopasowywanie wartości. Upraszczając, co bardziej wyrafinowane roboty trzeba by uczyć podstaw ludzkiej psychologii. Sztuczne urządzenia, a w szczególności autonomiczne pojazdy, muszą potrafić przewidywać ludzkie reakcje. Znaczy to, że jedni i drudzy muszą się uczyć, kim są ich potencjalni partnerzy. Jak zauważa Aga Dragan (2020), w procesie „uzgadniania wartości" trzeba przede wszystkim nauczyć roboty, że ludzie są czymś więcej niż „obiektami w ich środowisku". Wydaje się, że jest to bardzo ważne, a zarazem bardzo trudne. Dla człowieka widok piłki na środku jezdni i znajdującego się w pobliżu dziecka oznacza konieczność zachowania szczególnej ostrożności, gdyż jest całkiem prawdopodobne, że dziecko wbiegnie na jezdnię. Dla sztucznego systemu prawdopodobnie będą to dwa różne obiekty, które trzeba ominąć. Znaczy to, że roboty trzeba by wyposażać w coś, co w psychologii rozwojowej nazywa się teorią umysłu.

Alison Gopnik, wybitna przedstawicielka psychologii rozwojowej i twórczyni powyższej teorii, zafascynowana jest efektywnością uczenia się przedszkolaków. Z ich środowiska stara się czerpać inspiracje do uczenia inteligentnych agentów. Pisze ona: „Obserwacja dokonań dzieci może jednak dać programistom wskazówki dotyczące kierunków rozwoju uczenia komputerów. Dwie cechy dziecięcego podejścia są szczególnie uderzające. Dzieci są aktywnymi uczniami – nie chłoną biernie danych jak AI. (…) Mają naturalną motywację do wyłuskiwania informacji z otaczającego ich świata za pomocą niekończących się zabaw i eksploracji. (…) Być może sposobem na bardziej realistyczne i szersze uczenie jest zaimplementowanie maszynom ciekawości i umożliwienie im aktywnej interakcji ze światem" (Gopnik, 2018, s. 249).

Pojawia się też pytanie o to, na ile złożone algorytmy, projektowane m.in. z myślą o prognozowaniu ludzkich zachowań – np. do przewidywania recydywy w przypadku osób, które dokonały wykroczenia lub złamały prawo – są indyferentne wobec wartości. Jest to o tyle istotne, że w niektórych krajach oszacowane przez algorytm prawdopodobieństwo ponownego popełnienia zabronionego czynu wpływa na decyzję sądu w sensie konieczności zastosowania aresztu lub wysokości kary. Okazuje się, że mimo braku takich

intencji ze strony projektantów, algorytmy bywają „oskarżane" o stronniczość, czyli że uwzględniając takie, a nie inne zmienne, dyskryminują pewną grupę osób. Jak to jest możliwe? Otóż zauważmy, że złożone algorytmy, zanim pokażą na wyjściu wyniki wyrażone w postaci prawdopodobieństw, „karmione" są ogromną ilością danych zwanych Big Data, których szczegółowo nikt nie analizuje. Dane te odnoszą się do zapisów zarejestrowanych w ciągu ostatnich kilkudziesięciu lat, a wykorzystywane są do oszacowania częstości zdarzeń przyszłych. Dodajmy, że złożone algorytmy powstające w oparciu o uczenie się „sieci neuronowych" funkcjonują na zasadzie „czarnej skrzynki" – znamy źródło danych i końcowe wyniki, lecz nie znamy logiki procesów, które do nich doprowadziły. Wracając do przykładu, mogłoby okazać się – hipotetycznie – że 75% przestępstw danego typu dokonywali w przeszłości obywatele niebieskoocy. Zatem mogą oni oczekiwać, że sugerujący się tymi danymi sąd potraktuje ich surowiej niż osoby o innym kolorze oczu. Jednak mogło się zdarzyć, że związku ze zmianami o charakterze demograficznym, społeczno-kulturowym czy ekonomicznym w ostatnim roku proporcja osób niebieskookich w relacji do pozostałych oskarżonych uległa odwróceniu. Jeśli jednak algorytm opiera się na danych pochodzących ze znacznie dłuższego okresu czasu, niebieskoocy mogą poczuć się dyskryminowani, a algorytm można by uznać za stronniczy.

Można też wypełnić ten schemat/wzorzec inną treścią. Wyobraźmy sobie firmę ubezpieczeniową, która wyznacza wysokość składki w oparciu o dane dotyczące śmiertelności osób cierpiących na chorobę X w ciągu ostatnich 50 lat. Mogło się zdarzyć, że rok temu pojawił się nowy lek lub nowa strategia leczenia, co radykalnie zmniejszyło śmiertelność tych pacjentów. Jeśli jednak z przyczyn technicznych algorytm nie uwzględnił tych najświeższych danych, można by uznać go za stronniczy, a daną grupę pacjentów za dyskryminowanych przez ubezpieczyciela.

A zatem czy algorytmy SI są neutralne wobec ludzkich wartości? Jak już wspomniano, przestają być neutralne, jeśli zostały wypracowane w oparciu o niekompletną lub częściowo zdezaktualizowaną bazę danych, a ich wyniki wpływają na decyzje dotyczące konkretnych ludzi. Idąc dalej, można by stwierdzić, że właściwie wszelkie ustalenia dotyczące ludzi, które są upubliczniane (stając się częścią rzeczywistości społecznej) przestają być całkiem neutralne, gdyż wpływają na zachowania odbiorców w określonym kierunku. Nawiasem mówiąc, słuchowisko zrealizowane w latach 30. ubiegłego wieku na podstawie *Wojny światów* George'a Wellsa wywołało panikę. Znane też jest zjawisko samosprawdzających się przepowiedni (*self fulfilling prophecy*), kiedy to własnymi oczekiwaniami nieświadomie, acz skutecznie prowokujemy rzeczywistość do realizacji określonego scenariusza.

Konflikty wartości i problem osobistej wolności

Atak na World Trade Center w 2001 roku sprawił, że wyraźnie ujawniła się sprzeczność (konflikt) pomiędzy dwiema wartościami: wolnością jednostki a bezpieczeństwem. Doprowadziło to do postępującego ograniczania podstawowych praw człowieka w imię zwiększania bezpieczeństwa. Okazało się, że można inwigilować korespondencję i w ogóle prywatność obywateli w imię walki z terroryzmem i nie tylko. SI okazała się bardzo przydatna dla tych celów, a cyberprzestrzeń stała się polem walki politycznej (i hybrydowej).

W skali jednostkowej na ekspansję technologii cyfrowej warto spojrzeć również przez pryzmat innego dylematu, tj. doraźnej wygody vs prywatności i swobody wyboru. Wydaje się, że dylemat ten jest niedoceniany lub wręcz niedostrzegany. Gerd Gigerenzer (2022) w niedawno wydanej monografii *How to Stay Smart in a Smart World* odwołuje się do przekonującej metafory, która trafnie oddaje sytuację zwykłego użytkownika korzystającego z ofert dostępnych we „wszystko mającej" Sieci. Wyobraźmy sobie miasteczko, w którym istnieje tylko jedna kawiarnia i w dodatku serwują tam darmową kawę. Jest tylko drobna niedogodność – w kawiarni tej non stop nadawane są reklamy, a po sali krążą „domokrążcy", którzy namawiają klientów do zakupu różnych rzeczy (usług), zapewne po promocyjnych cenach. Wspomniany autor zauważa, że raz po raz, korzystając z darmowych niejako usług, bardzo chętnie sprzedajemy swoją prywatność. *De facto* mamy bowiem do czynienia z łagodnym szantażem – możesz skorzystać

z zawartości portalu lub zainstalować aplikację pod warunkiem ujawnienia swojego e-maila, numeru telefonu i innych danych personalnych. A w ramach bonusu możesz uzyskać zniżkę na zakupy lub bezpłatny newsletter.

Gigerenzer sugeruje, że rozsądniej i uczciwiej byłoby ponosić niewygórowane opłaty za te usługi niż opłacać je swoją prywatnością, która niechybnie zostanie wykorzystana co najmniej do celów komercyjnych. Okazuje się jednak, że istnieje swego rodzaju paradoks prywatności. Badania tegoż autora – przeprowadzone w Niemczech w 2019 roku na reprezentatywnej grupie 3200 osób w wieku powyżej 18 lat – pokazują, że wprawdzie 51% badanych uważa za największe związane z rewolucją cyfrową zagrożenie utratę prywatności i wynikającą stąd dostępność danych dla instytucji komercyjnych oraz rządowych. Równocześnie jednak na pytanie, czy byliby skłonni ponosić jakieś opłaty za ochronę ich danych w mediach społecznościowych, 75% zadeklarowało niechęć do ponoszenia jakichkolwiek kosztów. Tylko 18% uznało, że mogliby wnosić opłaty do 5 euro miesięcznie (Gigerenzer, 2022, s. 164). W innych, międzynarodowych badaniach – w których uczestniczyło 16 tysięcy osób w wieku powyżej 18 lat – tylko około 20% Europejczyków było skłonnych płacić 1$ miesięcznie za ochronę swych danych. Natomiast w takich krajach jak Zjednoczone Emiraty Arabskie, Brazylia, Meksyk i Chiny wskaźnik ten był o wiele wyższy i oscylował w okolicach 50% (Gigerenzer, 2022, s. 165). Wyniki te rzucają nieco światła (lub raczej cienia) na specyfikę i nieprzewidywalność ludzkiej racjonalności.

Gdy mowa o wolności osobistej w kontekście ucyfrowienia i ekspansji wirtualnej rzeczywistości, należałoby wspomnieć, że jedną z opozycji pojęciowych wolności jest uzależnienie. Uzależnienie od Sieci, a ściślej – od różnych oferowanych w niej usług, jest od wielu lat problemem społecznym o zasięgu globalnym. Dotyczy on setek milionów użytkowników „schwytanych w Sieć"[1] za pomocą gier i innych aplikacji projektowanych zgodnie z procedurą wyrafinowanego warunkowania instrumentalnego, gdzie w sposób nieregularny nagradza się odpowiednią reakcję. Procedura ta gwarantuje trwałość wytworzonych nawyków i odporność na ich wygaszanie. Temat to zbyt obszerny, wymagający osobnego potraktowania.

W stronę „sztucznej inteligencji emocjonalnej"

Mogłoby się wydawać, że ostatnim bastionem obrony przed ekspansją SI jest inteligencja emocjonalna (IE). Jej koncepcja zaistniała w psychologii w latach 90-tych ubiegłego wieku i szybko okazała się bardzo przydatna, zwłaszcza w kontekście psychologii stosowanej i rozwoju osobistego. To pojemne pojęcie odnosi się do umiejętności rozpoznawania, rozumienia oraz zarządzania własnymi emocjami, a także rozpoznawania i wpływania na emocje innych ludzi. Można w niej wyróżnić kilka składników czy aspektów. Przede wszystkim jest wśród nich umiejętność rozpoznawania, nazywania i wyrażania własnych uczuć oraz rozumienia ich uwarunkowań. Analogicznie dotyczy ona również umiejętności rozpoznawania i nazywania cudzych uczuć oraz adekwatnego (empatycznego) reagowania na uczucia innych ludzi. Zakłada też umiejętność „zarządzania" własnymi emocjami w sensie ich kontrolowania, czyli takiego sterowania nimi, aby sprzyjały skutecznej realizacji celów, radzenia sobie ze stresem i podtrzymywania odpowiedniego poziomu motywacji mimo zakłóceń oraz frustrujących okoliczności. Jest to ważne zwłaszcza w kontekście realizacji długoterminowych celów. Odpowiedni poziom IE jest również warunkiem nawiązywania i podtrzymywania harmonijnych relacji z innymi ludźmi, gdyż warunkuje lepszą komunikację i radzenie sobie z potencjalnymi konfliktami. Można powiedzieć, że inteligencja emocjonalna idzie w parze lub przechodzi w inteligencję społeczną.

Wydaje się, że procesy emocjonalne są szczególnie trudne do modelowania w ramach technologii cyfrowej. Okazuje się jednak, że od kilkunastu lat podejmowane są wybiórcze próby symulowania niektórych aspektów szeroko rozumianej emocjonalności, w tym także inteligencji

[1] Nawiązuję tu do tytułu klasycznej pracy Kimberly S. Young pt. *Caught in the Net: How to Recognize the Signs of Internet Addiction – and a Winning Strategy for Recovery* (1998).

emocjonalnej. Co więcej, w literaturze funkcjonuje już termin (Schuller, Björn, Schuller, 2018) „sztuczna inteligencja emocjonalna" (*artificial emotional intelligence*), co brzmi trochę jak oksymoron. Funkcjonują też pokrewne terminy, takie jak *affective computing* (Picard, 1997) lub *emotion AI* (Pietikäinen, Silvén, 2021). Prowadzone w różnych ośrodkach badania koncentrują się głównie na stosunkowo łatwym do „ucyfrowienia" aspekcie inteligencji emocjonalnej, czyli na rozpoznawaniu uczuć (innych) ludzi w oparciu o makro- i mikroekspresje mięśni twarzy (*action units*). Rejestrowana i analizowana jest mimika twarzy towarzysząca siedmiu podstawowym emocjom, wyróżnionym dawno temu przez Paula Ekmana (por. Ekman, Friesen, Hager, 2002). Są to: złość, strach, wstręt, szczęście, smutek, zdziwienie i neutralny stan emocjonalny. Okazuje się, że w mimice twarzy można wyróżnić około 30 punktów krytycznych, które zmieniają się zależnie od przeżywanej emocji.

Szczególnie wdzięcznym obszarem do badań przy użyciu technologii cyfrowej są mikroruchy towarzyszące pojawiającym się uczuciom, które manifestują się niezwykle krótko, bo od 0,03 do 0,5 sekundy. Posiadają one dwie, przyjazne dla badaczy, właściwości. Po pierwsze, nie podlegają intencjonalnej kontroli osoby, która mimowolnie je wyraża. Po drugie, w odróżnieniu od uczestniczących w interakcji osób (dla których jest to trudne zadanie), za pomocą sztucznych systemów można je łatwo rejestrować i wszechstronnie analizować.

Jeśli idzie o ten aspekt inteligencji emocjonalnej, jakim jest rozpoznawanie (rozróżnianie) cudzych stanów emocjonalnych, SI dysponuje już znaczącymi osiągnięciami (Dhope, Neelagar, 2022; Alisawi, Yalcin, 2023). W warunkach laboratoryjnych trafność rozpoznań przekracza niekiedy nawet 90%. W warunkach naturalnych jest to, oczywiście, nieporównanie trudniejsze.

Podejmowane badania nie sprowadzają się jedynie do modelowania tego aspektu AI, co jest zadaniem stosunkowo łatwym. Podejmowane są próby, by „emocjonalne" moduły włączyć do samego procesu uczenia się maszyn. Chodzi m.in. o to, by, podobnie jak w przypadku człowieka, aktywujące się „emocje" mogły zmieniać strategie przetwarzania informacji oraz (przynajmniej częściowo) zastępować zewnętrzne wzmocnienia odpowiednich reakcji wzmocnieniami wewnętrznymi. Tak czy inaczej wielu autorów zaangażowanych w ten nurt badań, (Schuller, Björn, Schuller, 2018; Pietikäinen, Silvén, 2021) jest przekonanych, że włączenie problematyki emocji do projektowania systemów SI może zrewolucjonizować tę dziedzinę i doprowadzić do kolejnego przełomu.

Autorzy monografii *Challenges of Artificial Intelligence: from Machine Learning and Computer Vision to Emotional Intelligence* (Pietikäinen, Silvén, 2021) dostrzegają wiele praktycznych zastosowań automatycznej identyfikacji cudzych stanów emocjonalnych w oparciu o zmiany mimiczne. Widzą ich zastosowanie w procesie leczenia poprzez monitorowanie mimiki pacjentów; w procesie edukacji (choćby poprzez wyłapywanie mikroekspresji świadczących np. o braku zrozumienia odbieranych informacji); przy monitorowaniu ekspresji mimicznej klientów supermarketów w reakcji na oglądane produkty; w kontekście rozmów kwalifikacyjnych, a nawet w odniesieniu do zabawek i gier wideo. Możemy więc w niedługim czasie spodziewać się, że wchodząc na dany portal lub chcąc skorzystać z jakiejś aplikacji, będziemy musieć wyrazić zgodę na „kamerowanie" naszych reakcji. Zależnie od rodzaju reakcji, czyli zidentyfikowanego stanu emocjonalnego, będziemy zachęcani do zakupu danych produktów lub usług. Zauważmy, że zdalne rejestrowanie mimicznie wyrażanych uczuć będzie na usługach różnych instytucji, zwiększając tym samym poziom kontroli poszczególnych jednostek. Póki co nie przewiduje się bowiem produkcji „kieszonkowych wykrywaczy emocji" na użytek pojedynczych osób, które ewentualnie chciałyby usprawnić swą skuteczność komunikacji z innymi ludźmi.

Pewne nadzieje dotyczące ograniczania tego typu inwigilacji można wiązać z przegłosowanym przez Parlament Europejski 14 czerwca bieżącego roku projektem AI Act, który proponuje szereg restrykcji[2], m.in. wprowadzenie zakazu biometrycznej identyfikacji osób w szeroko rozumianych

[2] Artificial Intelligence Act, European Parliament, https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf (dostęp: 21.06.2023).

sytuacjach publicznych (por. Łabanowicz, Jarczyńska, 2023). Wiadomo jednak, że zmiany prawne z natury swej są przesunięte w czasie i nie nadążają za zmianami rzeczywistości społecznej. Dodatkowe wątpliwości wiążą się też ze skutecznym ich egzekwowaniem.

Uwagi końcowe

A zatem, wbrew wcześniejszym optymistycznym oczekiwaniom związanym z pojawieniem się Internetu, ekspansja rewolucji cyfrowej prowadzi do większej centralizacji władzy i innych zasobów. Nowe technologie nigdy nie były sprzymierzeńcem demokracji. Doraźna wygoda wygrywa z wolnością jednostki rozpatrywaną w nieco dłuższej perspektywie. Bezpośrednie spotkania oraz kontakty interpersonalne (*face to face*) przegrywają z telekomunikacją i wirtualną rzeczywistością. Wprawdzie w psychologii wiadomo nie od dzisiaj, że ograniczenie kontaktów osobistych z innymi ludźmi sprzyja depresji (o czym też można było się przekonać w kontekście niedawnej pandemii) lecz niełatwo o tym pamiętać na co dzień. Chcąc uzyskać odrobinę dystansu do tej „normalnej", już oswojonej rzeczywistości, warto może sięgnąć ponownie po lekturę *Nowego wspaniałego świata* Aldousa Huxleya.

Bibliografia

Alisawi M., Yalcin N. (2023). Real-time emotion recognition using deep learning methods: systematic review. *Intelligent Methods in Engineering Sciences*, 2(1), 5–21, https://doi.org/10.58190/imiens.2023.7 (dostęp: 30.06.2023).

Artificial Intelligence Act, European Parliament, https://www.europarl.europa.eu/doceo/document /TA-9-2023-0236_EN.pdf (dostęp: 21.06.2023).

Brockman J. (ed.) (2020). *Possible Minds. 25 Ways of Looking at AI*. New York: Penguin Books.

Dhope P., Neelagar M.B. (2022). Real-time emotion recognition from facial expressions using Artificial Intelligence. *2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. Vijayawada, 1–6, https://doi.org/10.1109/AISP53593.2022.9760654, https://ieeexplore.ieee.org/document/9760654 (dostęp: 9.07.2023).

Dragan A. (2020). Putting the human into AI equation. In: J. Brockman (ed.), *Possible Minds. 25 Ways of Looking at AI*. New York: Penguin Books, 134–142.

Ekman P., Friesen W.V., Hager J.C. (2002). *Facial Action Coding System: The Manual on CD ROM*. Salt Lake City: A Human Face.

Fitch W.T. (2016). Nano-intentionality. In: J. Brockman (ed.), *What to Think About Machines That Think*. New York: Harper Collins Publisher, 89–92.

Gigerenzer G. (2022). *How to Stay Smart in a Smart World. Why Human Intelligence Still Beats Algorithms*. London: Allen Lane.

Gopnik A. (2018). AI kontra czterolatek. W: J. Brockman (red.), *Człowiek na rozdrożu. Sztuczna inteligencja – 25 punktów widzenia*. Gliwice: Helion, 239–250.

Leijen I., van Herk H., Bardi A. (2022). Individual and generational value change in an adult population, a 12-year longitudinal panel study. *Scientific Reports*, 12:17844, https://doi.org/10.1038/s41598-022-22862-1, https://www.nature.com/srep/ (dostęp: 22.06.2023).

Łabanowicz K., Jarczyńska A. (2023). Projekt unijnego prawa w sprawie sztucznej inteligencji przyjęty. To będzie rewolucja na miarę RODO, https://www.onet.pl/biznes/forbes/parlament-europejski-przyjal-projekt-ai-act-sztuczna-inteligencja (dostęp: 14.06.2023).

Mayers D.G., Twenge J.M. (2018). *Social Psychology*, ed. XIII. New York: McGraw Hill.

Mudyń K. (2010). Digitalizacja rzeczywistości a problem dekontekstualizacji istnienia. W:T. Rowiński, R. Tadeusiewicz (red.), *Psychologia i informatyka. Ich synergia i kontradykcje*. Warszawa: Wydawnictwo UKSW, 191–204.

Mudyń K. (2012). O różnych aspektach antropomorfizacji, „systemach intencjonalnych" i dyskretnym uroku technologii. W: J. Morbitzer, E. Musiał (red.), *Człowiek–Media–Edukacja*. Kraków: Wydaw. KTiME UP, 307–312.

Mudyń K. (2014). Miedzy antropomorfizacją a dehumanizacją. Powracający problem natury ludzkiej. *Czasopismo Psychologiczne*, 1(20), 1–9.

Mudyń K. (2022). „Człowiek na rozdrożu. Sztuczna inteligencja – 25 punktów widzenia"– recenzja. *Tygodnik Spraw Obywatelskich*, 137(33), https://instytutsprawobywatelskich.pl/krzysztof-mudyn-czlowiek-na-rozdrozu-recenzja/ (dostęp: 25.07.2023).

Mudyń K. (2022). W poszukiwaniu biocentrycznej definicji inteligencji. Rośliny są inteligentniejsze od „inteligentnych maszyn", https://www.researchgate.net/publication/365568791_W_poszukiwaniu_biocentrycznej_definicja_inteligencji_Rosliny_sa_inteligentniejsze_od_inteligentnych_maszyn_In_search_of_a_biocentric_definition_of_intelligence_Plants_are_smarter_than_%27smart_machines%27 (dostęp: 1.03.2023).

Nisbett R.E. (2016). *Mindware. Narzędzia skutecznego myślenia*. Sopot: Wydawnictwo „Smak Słowa".

Picard R. (1997). *Affective Computing*. Cambridge, MA: MIT Press

Pietikäinen M., Silvén O. (2021). *Challenges of Artificial Intelligence: from Machine Learning and Computer Vision to Emotional Intelligence*, http://urn.fi/urn:isbn:9789526231990 (dostęp: 3.01.2023).

Rokeach M. (1973). *The Nature of Human Values*. New York: The Free Press.

Russell S. (2020). O przekazywaniu maszynom ogólnych celów. W: J. Brockman (red.), *Człowiek na rozdrożu. Sztuczna inteligencja – 25 punktów widzenia.* Gliwice: Wydawnictwo „Helion", 41–53.

Schuller D., Björn W., Schuller B.W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9), 38–46, DOI:10.1109/MC.2018.3620963, https://ieeexplore.ieee.org/document/8481266 (dostęp: 5.07.2023).

Young K.S. (1998). *Caught in the Net: How to Recognize the Signs of Internet Addiction – and a Winning Strategy for Recovery.* New York: John Wiley & Sons.

Walsh T. (2018). *To żyje. Sztuczna inteligencja. Od logicznego fortepianu po zabójcze roboty.* Warszawa: Wydawnictwo Naukowe PWN.

Wiener N. (1950). *The Human Use of Human Beings. Cybernetics and Society*, https://monoskop.org/images/6/60/Wiener_Norbert_The_Human_Use_of_Human_Beings_1989.pdf (dostęp: 9.08.2021).

Wiener N. (1961). *Cybernetyka i społeczeństwo*. Warszawa: Wydawnictwo KiW.

Źródła finansowania / Funding sources: brak źródeł finansowania / no sources of financing
Konflikt interesów / Conflict of interest: brak konfliktu / no conflict of interest
Otrzymano/Received: 12.07.2023
Zaakceptowano/Accepted: 30.07.2023